DOCUMENT RESUME

ED 445 059 TM 031 675

AUTHOR Plake, Barbara S.; Buckendahl, Chad W.; Impara, James C.
TITLE A Comparison of Publishers' and Teachers' Perspectives on the Alignment of Norm-Referenced Tests to Nebraska's Language Arts Content Standards.
PUB DATE 2000-06-00
NOTE 15p.; Paper presented at the Annual National Conference of the Council of Chief State School Officers on Large-Scale Assessment (30th, Snowbird, UT, June 25-28, 2000).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; *Language Arts; *Norm Referenced Tests; *State Standards; Student Evaluation; Teacher Attitudes; *Teachers; Test Use
IDENTIFIERS *Nebraska; *Test Publishers

ABSTRACT

Test publishers have promoted their commercially available norm-referenced achievement tests as viable solutions to assessment challenges faced by states. They argue that their tests are developed professionally and therefore possess sound psychometric properties not often found in state-specific efforts. This study compared judgments from two sources, test publishers and teachers, on the alignment of test items from two commercially available norm-referenced achievement tests to Nebraska's content standards at three grade levels (4, 8, and 11) in language arts. Study analyses focused on the level of agreement between the state's teachers' perceptions of how well the tests aligned to the standards and the alignment reported by the two test publishers. Panels of 20 4th grade teachers, 10 8th grade teachers, and 10 11th grade teachers studied the alignment. Results indicate that there may be some inconsistency between teachers' and publishers' perceptions. (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 445 059

A comparison of publishers' and teachers' perspectives on the alignment of norm-referenced tests to Nebraska's language arts content standards

Barbara S. Plake

Chad W. Buckendahl

James C. Impara

Buros Center for Testing

Paper presented at the annual Large Scale

Assessment Conference in Snowbird, Utah

June, 2000

TM031675

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

C. W. Buckendahl

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

Abstract

Test publishers have promoted their commercially available, norm-referenced achievement tests as viable solutions to assessment challenges faced by states. They argue that their tests are developed professionally, and therefore possess sound psychometric properties not often found in state-specific efforts. This study compared judgments from two sources, test publishers and teachers, on the alignment of test items from two commercially available norm-referenced achievement tests to Nebraska's content standards at three grade levels (4^{th}, 8^{th}, and 11^{th}) in language arts. Analyses in the study focused on the level of agreement between the state's teachers' perceptions of how well the tests aligned to the standards and the alignment reported by the two test publishers. Results indicate that there may be some inconsistency between teachers' and publishers' perceptions.

A comparison of publishers' and teachers' perspectives on the alignment of norm-referenced tests to Nebraska's language arts content standards

Introduction

As part of the educational reform movement, many states have identified content-specific standards for student achievement at various grade levels. Frequently following the articulation of these content standards is an interest in assessing how well students perform relative to these content standards. Decisions on appropriate assessment strategies that adequately measure students' performance on these standards represent a substantial challenge (Linn & Herman, 1997).

Faced with these difficult decisions, it is not surprising that some states have struggled to find adequate solutions. One option for states may be to select a commercially available, norm-referenced achievement test that would report student performance relative to national norms but provide less state-specific information. A second option for states may be to contract for the development of a criterion-referenced state test. This second option would likely provide a better match between the test and the state's content standards. More importantly, it would likely provide better validity evidence for resultant scores. The trade-off, however, would be to impose a heavier financial and resource burden on the state than would the purchase of a commercially available test.

Alignment studies that examine the relationship between test items and criteria, generally provide some of the validity evidence needed to make inferences about tests' scores as they relate to a set of objectives (Webb & Smithson, 1999; Baker, Freeman, &

Clayton, 1991). A state's adopted content standards are typically used as the objectives to which tests are aligned. Since decisions may be made about student or district performance based on the results of the test, a close match between what is expected and what is measured is crucial (Webb, 1997). The issue then becomes which individuals, groups, or organizations can best provide the judgments for this important alignment information.

Test publishers have promoted their commercially available, norm-referenced achievement tests as viable solutions to these assessment challenges faced by states. They argue that their tests are developed professionally, and therefore possess sound psychometric properties not found in state-specific efforts. Moreover, they specify that the curricular coverage provided by their tests often adequately represents the content standards of the state. To provide evidence of the alignment of their tests to a state's content standards (and to market their product), many test publishers produce documents that show how their tests are aligned with the state content standards (Harcourt Brace Educational Measurement, 1999; Riverside Publishing, 1998).

The purpose of this study was to compare judgments from publishers and teachers on the alignment of test items from two commercially available norm-referenced achievement tests to Nebraska's content standards in language arts. The goal of the study was to determine the level of agreement between test publishers and teachers when examining their judgments on alignment of test items to the state's content standards.

Methods

This study considered the content standards for Language Arts (reading, writing, speaking, and listening) at the fourth, eighth and high school levels in Nebraska. In

separate validity studies for each grade level, panels of experienced practicing teachers aligned test questions from the five most prevalent achievement test batteries used in Nebraska to the language arts content standards. In addition to these validity studies, two of the test publishers had independently conducted an alignment of their achievement test batteries to the state's content standards. Analyses in our study focused on the level of agreement between the teachers' perceptions of how well the tests aligned to the standards and the alignment reported by the test publishers.

Procedure for Teachers' Alignment

The same basic procedure was followed for the validity studies at each of the three grade levels. Teachers were identified by their school district to participate in this project. In order to represent the state as much as possible, the state was sectioned into 10 geographic areas. Districts were contacted within these areas to identify teachers experienced at the grade level and content area to participate in the project, resulting in 10 teachers on each of these panels. For the fourth grade panel, this number was increased to 20 teachers because teachers at this grade are generally responsible for more than just one content area. These teachers came to a central location to participate in the 2-day validity study for their grade level.

In advance of the meeting, teachers received basic information about the project and their participation. They were also sent a copy of the content standards for their specific grade and content area and asked to review them thoroughly before attending the project meeting. These content standards had been only recently adopted statewide and had received substantial attention from both educators and media. Therefore, these

participating teachers were likely knowledgeable in these standards; nonetheless, they were instructed to familiarize themselves with the standards prior to arrival.

Upon arrival, teachers were given a brief general orientation. Teachers first reviewed the content standards for the grade level and clarifications of interpretations of the language arts standards were provided. The teachers also participated in an activity designed to provide additional familiarization with their grade and content standards. Next, the teachers were trained in the rating process they would use to provide their judgments for the alignment of the standardized achievement test items to the content standards.

Teachers used the following rating criteria to judge the level of alignment of an item to standard:

High Level of Alignment = A high level of alignment indicates that the item measures the standard. This high rating means that you would be very comfortable making inferences about a student's performance on that standard by knowing their performance on this item or similar items.

Moderate Level of Alignment = A moderate level of alignment indicates that the item measures a portion of the standard. Since standards may have multiple parts, a high level of alignment may not be appropriate, but portions that are addressed by an item may warrant this level of alignment. A moderate rating means that you would be somewhat comfortable making inferences about a student's performance on that standard by knowing their performance on a collection of similar items.

Low Level of Alignment = A low level of alignment indicates that the item barely measures the standard. This level of alignment means that you probably had to stretch to

find alignment between item and standard. An item that aligns with only a small portion of a standard comprised of many aspects may warrant this rating. A low rating means that you would feel less than somewhat comfortable making inferences about a student's performance on the standard knowing their performance on this item or similar items.

No Alignment = No alignment means that an item does not measure any aspects of the standard. This lack of alignment means that you found no alignment between item and standard. This rating means that you would not make an inference about a student's performance on the standard knowing their performance on this item or similar items.

As a group, the teachers rated and discussed several sample items to familiarize themselves with the types of test items they would be rating, the rating form, and the rating scale. They were instructed to use the definitions above with the following rating scale to record their judgments: "H" = high alignment; "M" = moderate alignment; "L" = low alignment and "blank" = no alignment. Teachers were instructed to identify all matches of a test item to the standards; items that had multiple potential matches to the standards were modeled in the practice[1]. Following training, the teachers were given the test booklets and forms and were asked to independently evaluate the alignment of the test questions to the state content standards. The order of the test booklets was balanced so that half of the teachers evaluated the item-to-standards match using one order and the other half considered the test booklets in the reversed order. Within each test battery, only the sub-tests identified for the "Complete" battery were considered; therefore, none of the additional or supplemental test materials were included in the project.

[1] Because the standards are rather broadly stated, some items could be used to make inferences about performance on more than one standard.

Analysis

The results of this study are predicated on the operational definition of adequate measurement of a standard that was used in this study. There were three criteria used to determine whether a standard was adequately measured: 1) alignment rating level, 2) teacher agreement on rating levels, and 3) number of aligned items. Item-to-standard matches that were identified by teachers who used the high or moderate alignment levels met the first criterion for adequate measurement of a standard. If at least 50% of teachers (5 for 8th grade and high school panels, 10 for 4th grade panels) agreed with this high or moderate level of alignment, the second criterion for adequate measurement of a standard was met. Last, there had to be at least five items that met criteria 1) and 2). Thus, a standard was said to be measured if a minimum of five items a) were rated at the high or moderate level of alignment by b) at least 50%. Although Webb (1999) suggests a minimum of six items for making an inference about content knowledge, we chose five items as a more lenient criterion since there was a consideration of inter-rater reliability criterion (Schmidt, 1999) included as well. These results were tabulated for the Language Arts content standards for each of the three grade levels for the five most prevalent standardized achievement tests in the state.

Comparison to Publisher's Alignments

Publishers for two of the achievement tests used in this project independently prepared materials identifying the alignment of their achievement test batteries to the state's content standards. Although these publishers frequently considered all relevant sub-tests in the total battery, the comparison was only based on the rating of the sub-tests within the "complete" battery. Therefore, the results from the teachers' alignment

judgments and those by the publishers were directly comparable. Riverside Publishing (1998) used "educators who have extensive knowledge of current curriculum and learning theory, and who are familiar with Riverside's products (p. vi)." However, the exact procedures used by these two publishers were not reported. Their reports of items-to-standards match were the only information used for the comparisons. The criterion used to compare the relative agreement of the publishers with the teachers was based on the number of standards each group (teachers or publishers) indicated were adequately measured. For the most part, publishers generally reported that there were 2-3 items that measured each standard in this study, the criterion for adequate measurement of a standard was set at a minimum of five items to remain consistent with the number of items required in the teachers' criteria.

Comparative Analyses

In analyzing the comparison between publishers' judgments of alignment to the standards and teachers' judgment of alignment, a decision rule was formulated that looked at the each content area by standard. Standards judged to be adequately measured by either teachers or publishers were classified into four distinct categories. The first category we called, "Teacher Only," meaning only teachers, not publishers found adequate alignment of items to a content standard. The second category was "Publisher Only," meaning only publishers, not teachers found adequate alignment of items to a content standard. The third category, "Teachers ≥ Publishers" indicates that both teachers and publishers ratings met the criteria for adequate alignment of items to a content standard, but teachers judged that an equal or greater number of items across the sub-tests measured the standard than did publishers. The fourth category, "Publishers >

Teachers," also means that ratings from both groups met the criteria for adequate alignment of items to a content standard, but publishers reported that more items aligned to the standard than did teachers. In the results, these categories are presented as frequencies and as a ratio of teacher-publisher agreement for standards judged to be adequately measured.

Results

Results show inconsistencies between the publishers' perceptions and those of practicing teachers regarding the alignment of standardized achievement tests and the state's content standards. For example, in fourth grade Language Arts, teachers found matches for six of the sixteen standards (38%) for both Tests A and B, whereas Publisher A found matches for eleven of the sixteen standards (69%) and Publisher B found matches for fourteen of the sixteen standards (88%). Table 1 shows the breakdown by category of teachers' and publishers' judgments for the two tests for which publishers submitted alignment reports.

TABLE 1. Grade 4 Language Arts Standards[2] Judged Adequately Aligned by Test.

Category	Test A	Test B
Teacher Only	0	0
Publisher Only	5	8
Teacher ≥ Publisher	4	3
Publisher > Teacher	2	3
Teacher/Publisher Agreement	6/11 (55%)	6/14 (43%)

[2] Based on 16 language arts standards at this grade level

There was an overlap of six of the sixteen standards (38%) for both Tests A and B indicated by both teachers and publishers. For Publisher A, the agreement with teachers was on six of eleven (55%) standards judged to be adequately measured by one group or both. For Publisher B, the agreement with teachers was on only six of fourteen (43%) standards.

For eighth grade Language Arts, teachers found matches for seven of the sixteen standards (44%) for both Tests A and B, whereas Publisher A found matches for eleven of the sixteen standards (69%) and Publisher B found matches for thirteen of the sixteen standards (81%). Table 2 shows the breakdown by category of teachers' and publishers' judgments for the two tests for which publishers submitted alignment reports.

TABLE 2. Grade 8 Language Arts Standards[3] Judged Adequately Aligned by Test.

Category	Test A	Test B
Teacher Only	2	0
Publisher Only	6	6
Teacher ≥ Publisher	2	4
Publisher > Teacher	3	3
Teacher/Publisher Agreement	5/13 (38%)	7/13 (54%)

There was an overlap of five of the sixteen standards (31%) and seven of the sixteen standards (44%) for Tests A and B respectively, indicated by both teachers and publishers. For Publisher A, the agreement with teachers was on five of thirteen (38%)

[3] Based on 16 language arts standards at this grade level

standards judged to be adequately measured by one group or both. For Publisher B, the agreement with teachers was on seven of thirteen (54%) standards.

For high school Language Arts, teachers found matches for six of the sixteen standards (44%) and eight of the sixteen standards (50%) for Tests A and B, respectively. Conversely, Publisher A found matches for eleven of the sixteen standards (69%) and Publisher B found matches for fifteen of the sixteen standards (94%). Table 3 shows the breakdown by category of teachers' and publishers' judgments for the two tests for which publishers submitted alignment reports.

TABLE 3. High School Language Arts Standards[4] Judged Adequately Aligned by Test.

Category	Test A	Test B
Teacher Only	0	0
Publisher Only	5	7
Teacher ≥ Publisher	4	3
Publisher > Teacher	2	5
Teacher/Publisher Agreement	6/11 (55%)	8/15 (53%)

There was an overlap of six of the sixteen standards (38%) and five of the sixteen standards (31%) for Tests A and B respectively, indicated by both teachers and publishers. For Publisher A, the agreement with teachers was on six of eleven (55%) standards judged to be adequately measured by one group or both. For Publisher B, the agreement with teachers was on eight of fifteen (53%) standards.

[4] Based on 16 language arts standards at this grade level

Implications for State Assessment Programs

Test publishers have a vested interest in maximizing the alignment of their standardized achievement test batteries to the state's content standards. In Nebraska, these publishers have been promoting the validity of their tests as adequately measuring the state's content standards, pointing to the results of their validity/alignment analyses. Based on the results of this study, caution should be used in evaluating these publishers' statements regarding the alignment of their achievement tests to a state's content standards. Further, states are encouraged to conduct their own validity studies prior to making a decision regarding an assessment. Results indicate that for Language Arts, publishers frequently envisioned matches of items to content standards that were not endorsed by teacher panels. An additional finding is that these publishers base their assessment alignments on fewer items than may be preferable for making reasonable inferences regarding student performance. This may be an important consideration for state or district assessment programs that use student scores from these tests for any high stakes decisions regarding rewards or sanctions.

References

Baker, E.L., Freeman, M., & Clayton, S. (1991). Cognitive assessment of history for large-scale testing. In M.C. Wittrock & E.L. Baker (Eds.), *Testing and cognition* (pp. 131-153). Englewood Cliffs, NJ: Prentice Hall.

Harcourt Brace Educational Measurement (1999). Nebraska Content Standards Match to *Stanford Achievement Test* Series, Ninth Edition. San Antonio, TX: Harcourt Brace Educational Measurement.

Linn, R.L & Herman, J.L. (1997). *A Policymaker's Guide to Standards-Led Assessment*. Denver, CO: Education Commission of the States & National Center for Research on Evaluation, Standards, and Student Testing.

Riverside Publishing (1998). Correlation of Nebraska Academic Standards to the Iowa Tests. Itasca, IL: Riverside Publishing.

Schmidt, W.H. (1999). *Using off-the-shelf tests in large-scale assessment systems*. Presentation at the National Conference on Large-Scale Assessment. Snowbird, UT.

Webb, N.L. & Smithson, J. (1999, April). *Alignment between standards and assessments in mathematics for grade 8 in one state*. Paper presented at the annual meeting of the American Educational Research Association. Montreal, Quebec.

Webb, N.L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education*. Council of Chief State School Officers and National Institute for Science Education Research Monograph No. 6. Madison, WI: University of Wisconsin – Madison.

Wixson, K.K., Fisk, M.C., Dutro, E., & McDaniel, J. (1999, June). *The alignment of state standards and assessments in elementary reading*. Presentation at the National Conference on Large-Scale Assessment. Snowbird, UT.